BlueGene/L Supercomputer Hardware

Gerard V. Kopcsay

IBM Research

October 14, 2003

BlueGene/L Features

- Scalable – from half rack to hundreds of racks
- One to two orders of magnitude improvement in
  - Peak performance,
  - Price performance,
  - Floor space per Teraflop/s,
  - Power per Teraflop/s.
- High packaging density – 1024 compute nodes per rack
  - Enabled by low power, system-on-a-chip ASIC technology.
  - Use standard proven components wherever possible to improve reliability and reduce cost.
  - Design advanced components where needed for increased application performance.
  - Develop air cooled rack configuration for up to 25 kW power.

Cost/Performance

- BlueGene/L is cost/performance optimized for a wide class of parallel applications.

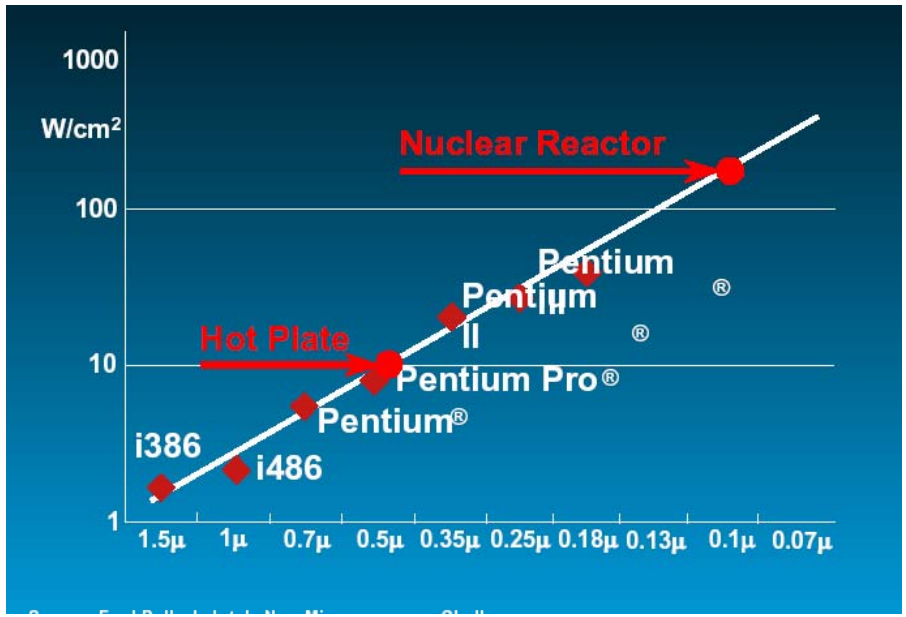
Cost

- Machine
- Facilities
- Hardware Support and Maintenance
- Software Support
  - system
  - application

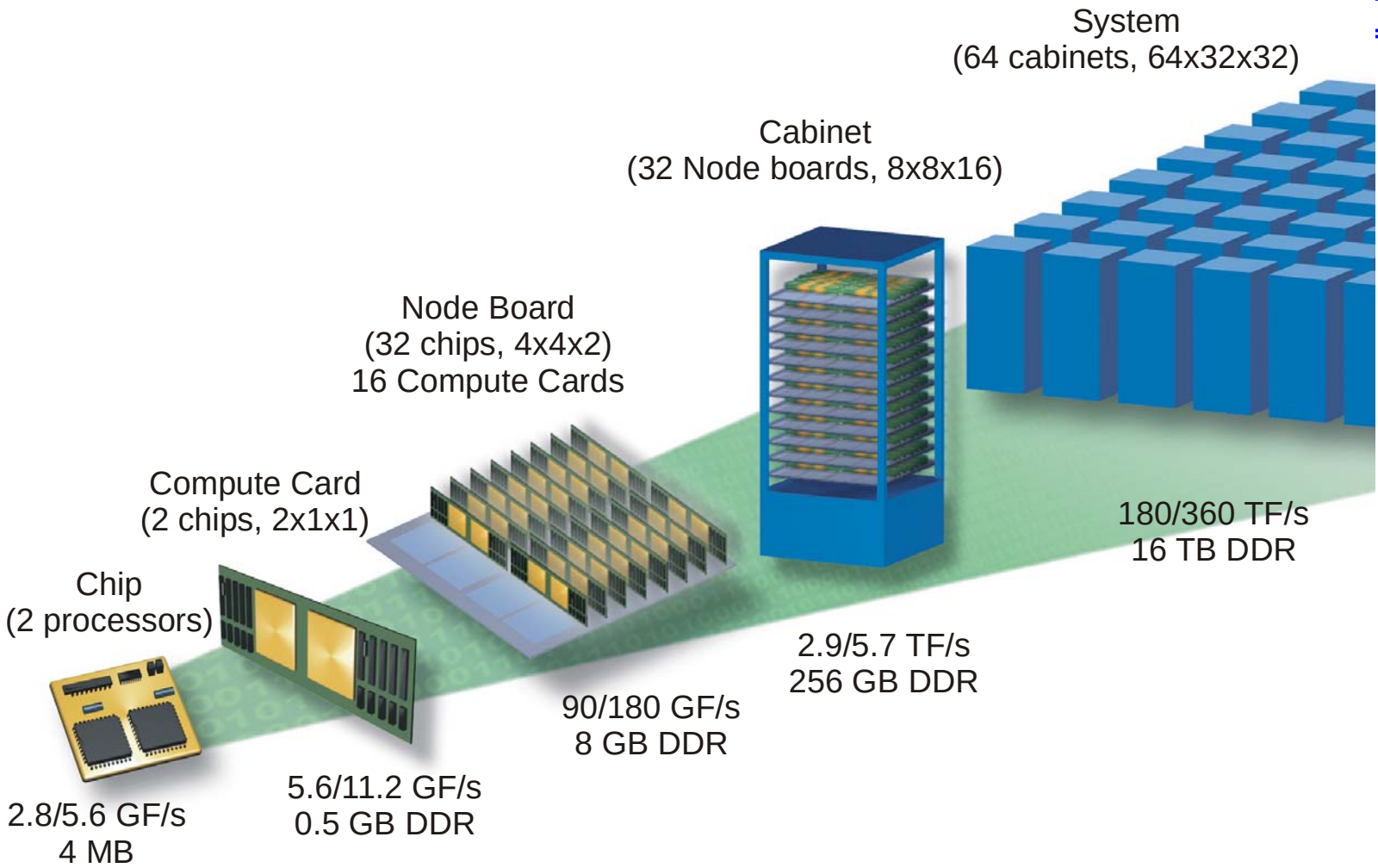
power is the dominant factor

Performance

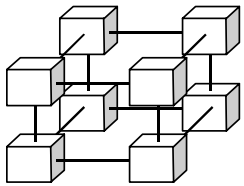
- Peak speed
- Scaleability
- Availability
- Useability
  - tools , debuggers, performance analysis
  - compilers, libraries, frameworks



BlueGene/L

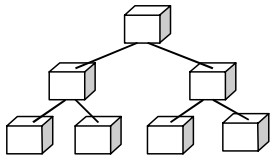

System
(64 cabinets, 64x32x32)
Cabinet
(32 Node boards, 8x8x16)
Node Board
(32 chips, 4x4x2)
16 Compute Cards
Compute Card
(2 chips, 2x1x1)
Chip
(2 processors)
2.8/5.6 GF/s
4 MB
5.6/11.2 GF/s
0.5 GB DDR
90/180 GF/s
8 GB DDR
2.9/5.7 TF/s
256 GB DDR
180/360 TF/s
16 TB DDR

The BlueGene/L Networks



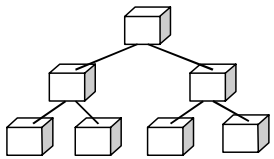
3 Dimensional Torus

- Point-to-point



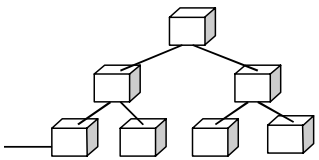
Global Tree

- Global Operations



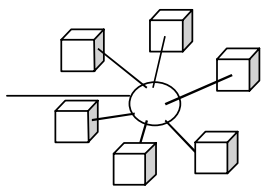
Global Barriers and Interrupts

- Low Latency Barriers and Interrupts



Gbit Ethernet

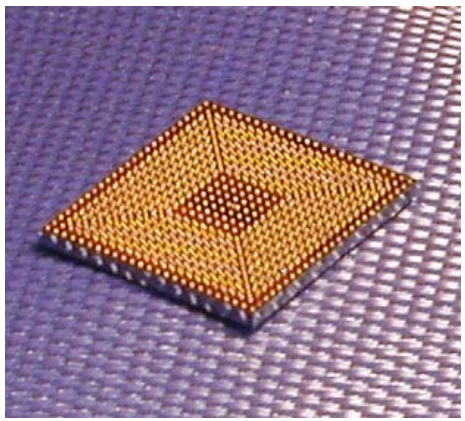
- File I/O and Host Interface



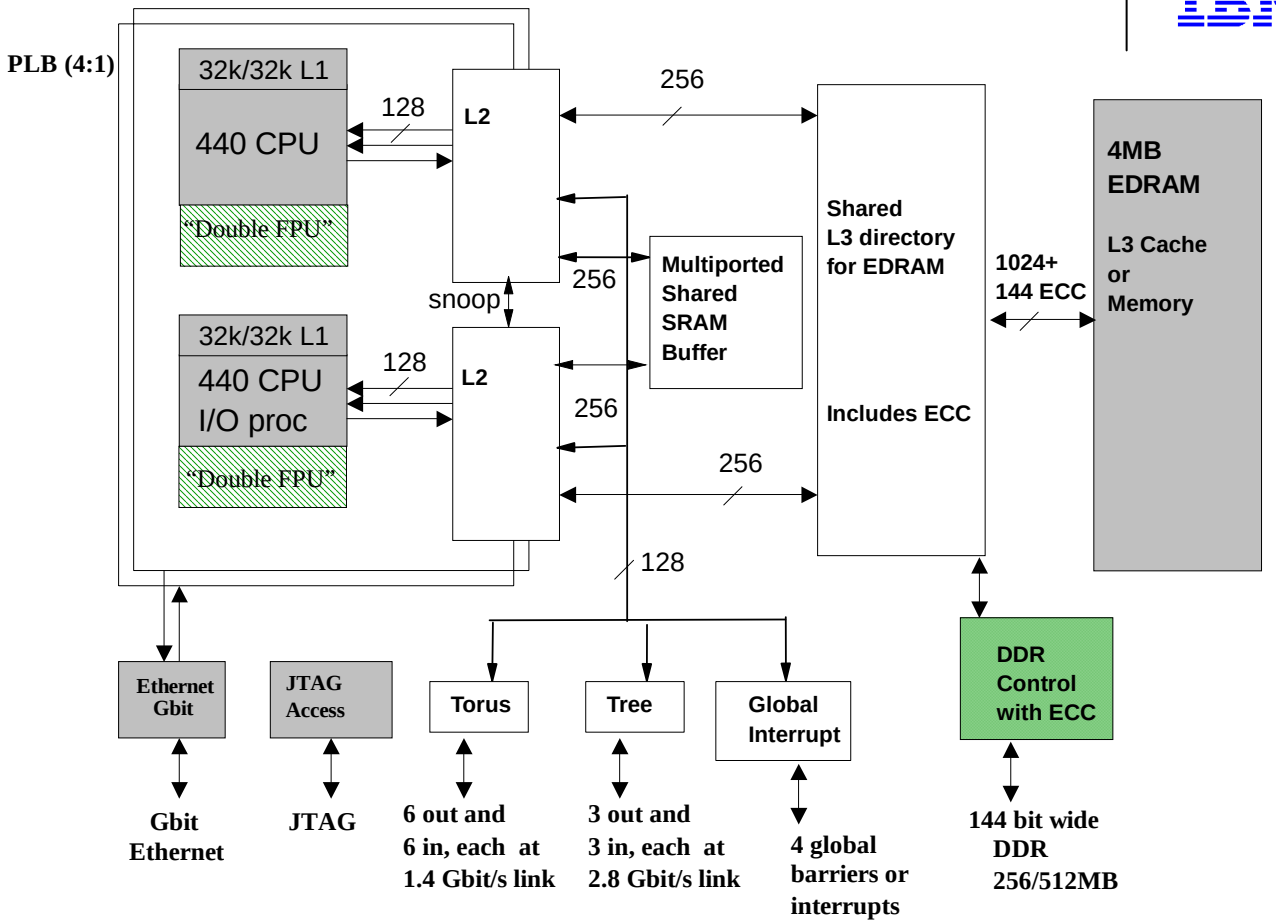
Control Network

- Boot, Monitoring and Diagnostics

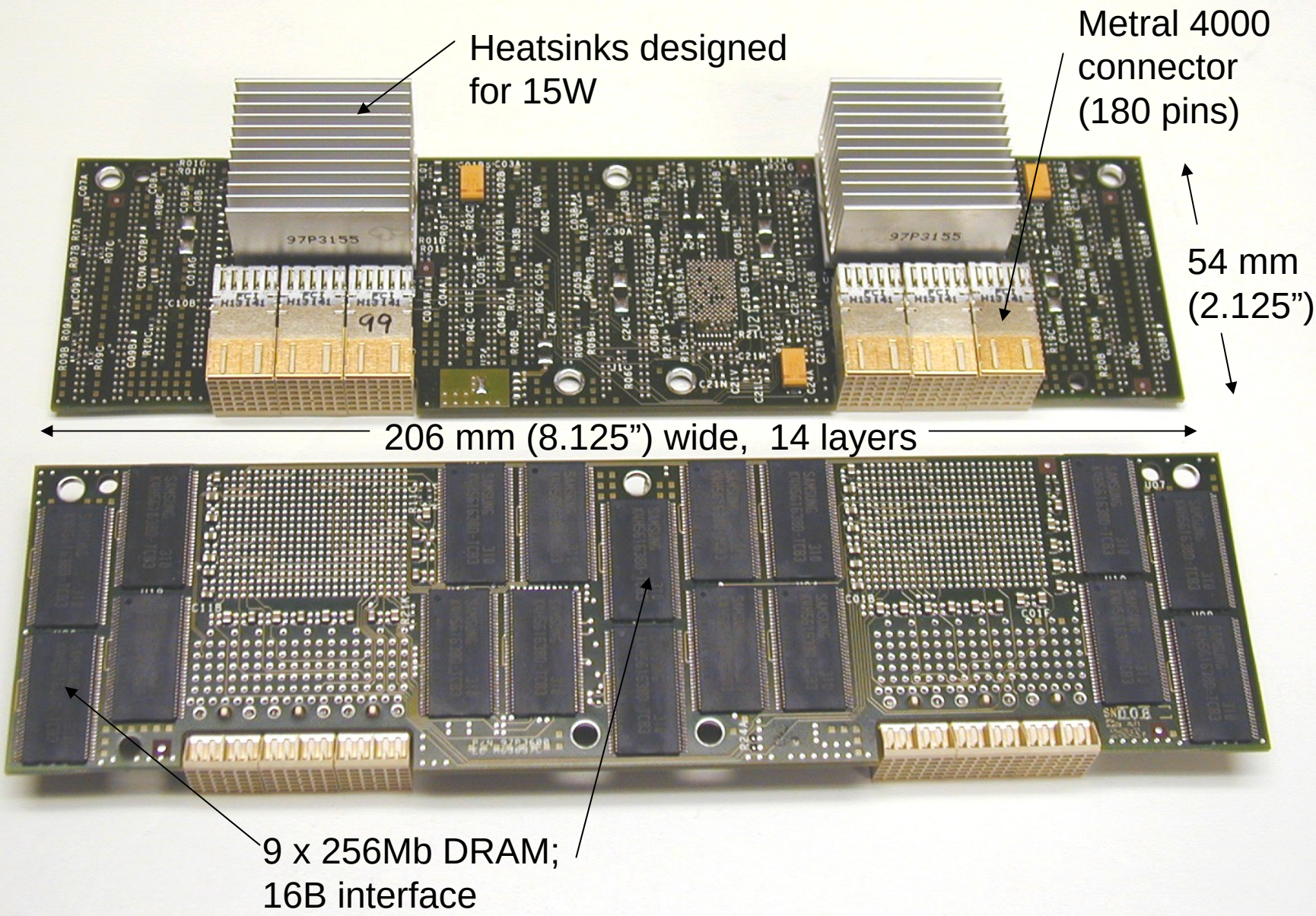
BlueGene/L Compute ASIC

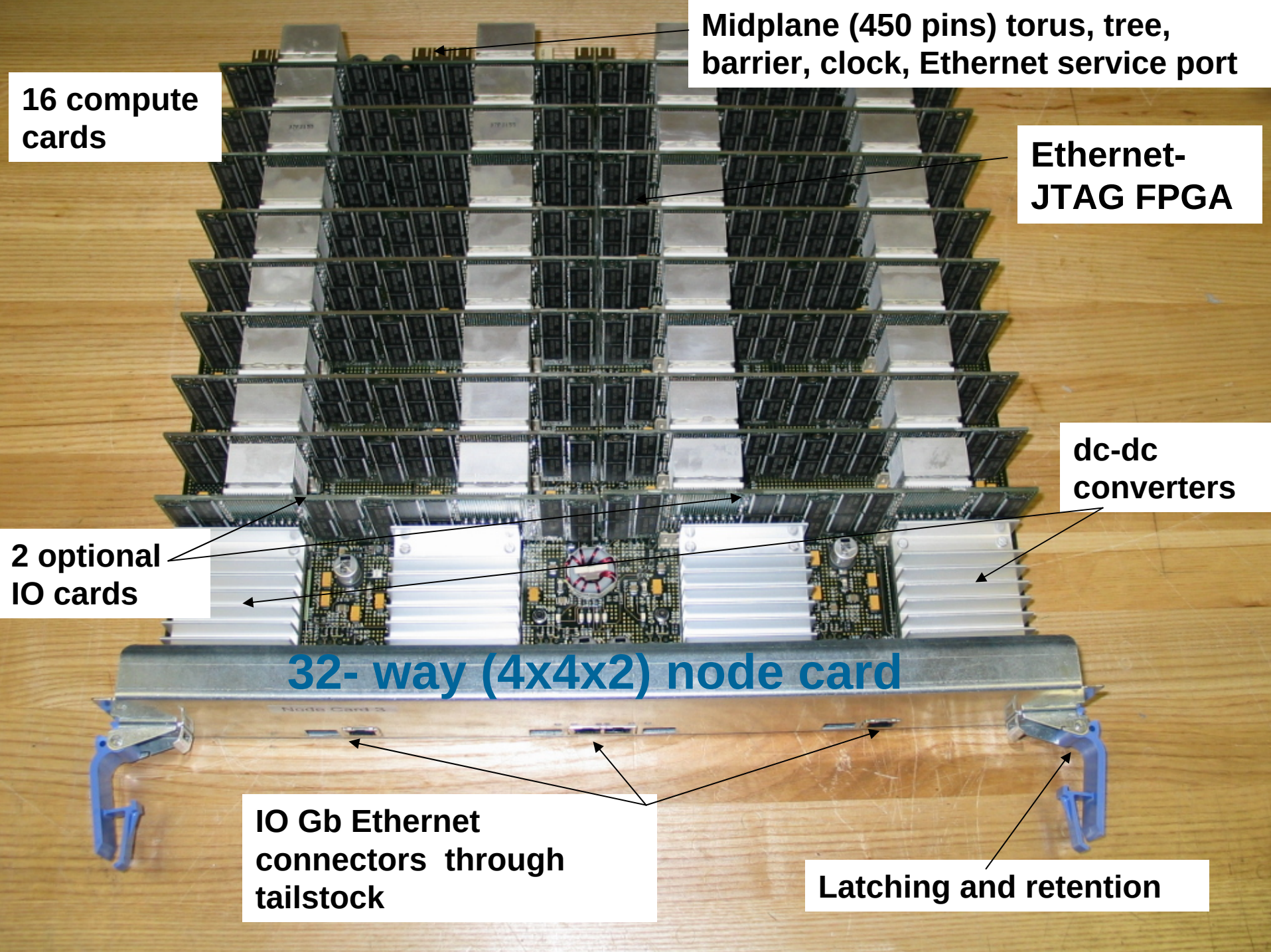


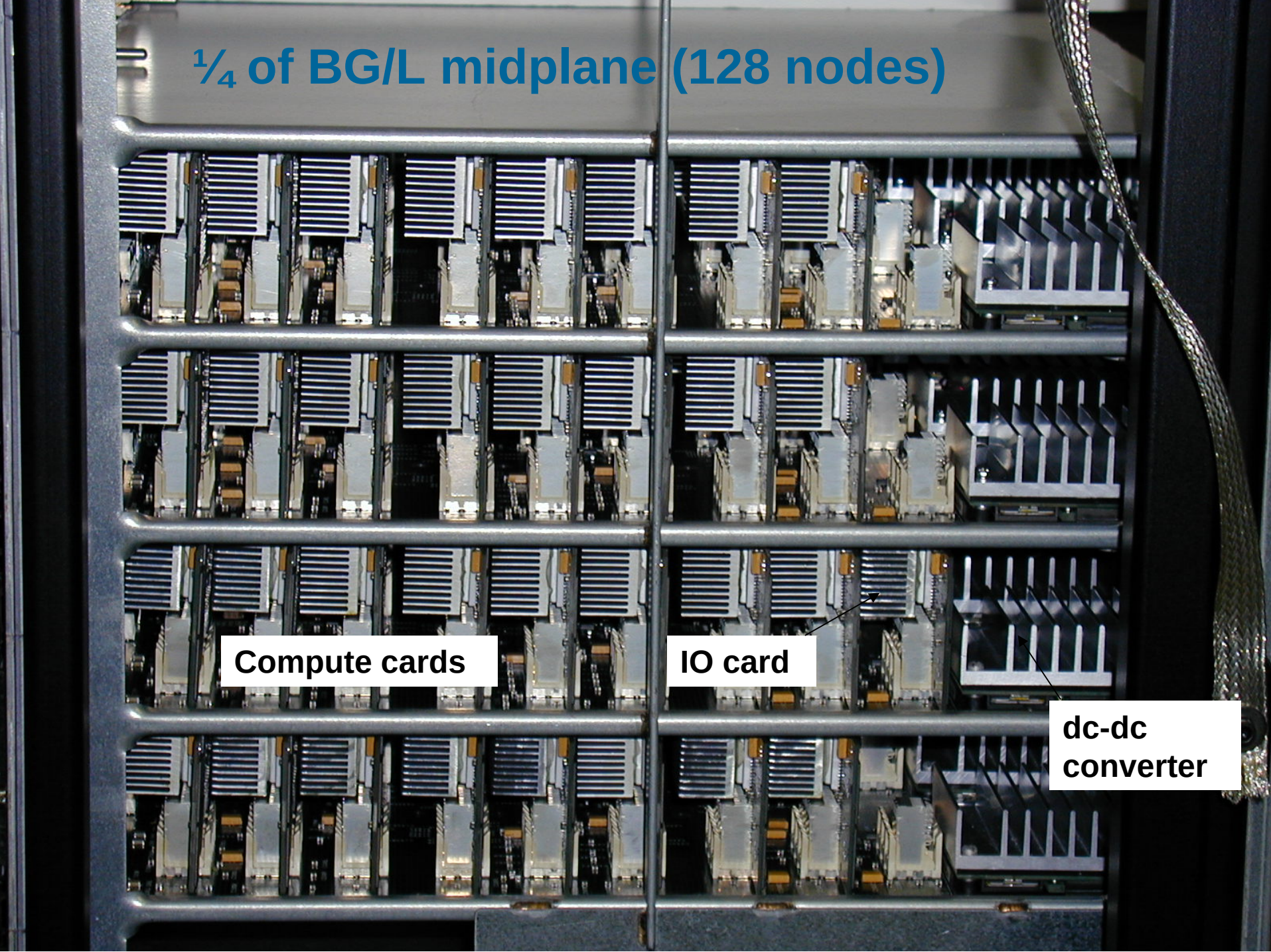
- IBM CU-11, 0.13 µm
- 11 x 11 mm die size
- 25 x 32 mm CBGA
- 474 pins, 328 signal
- 1.5/2.5 Volt



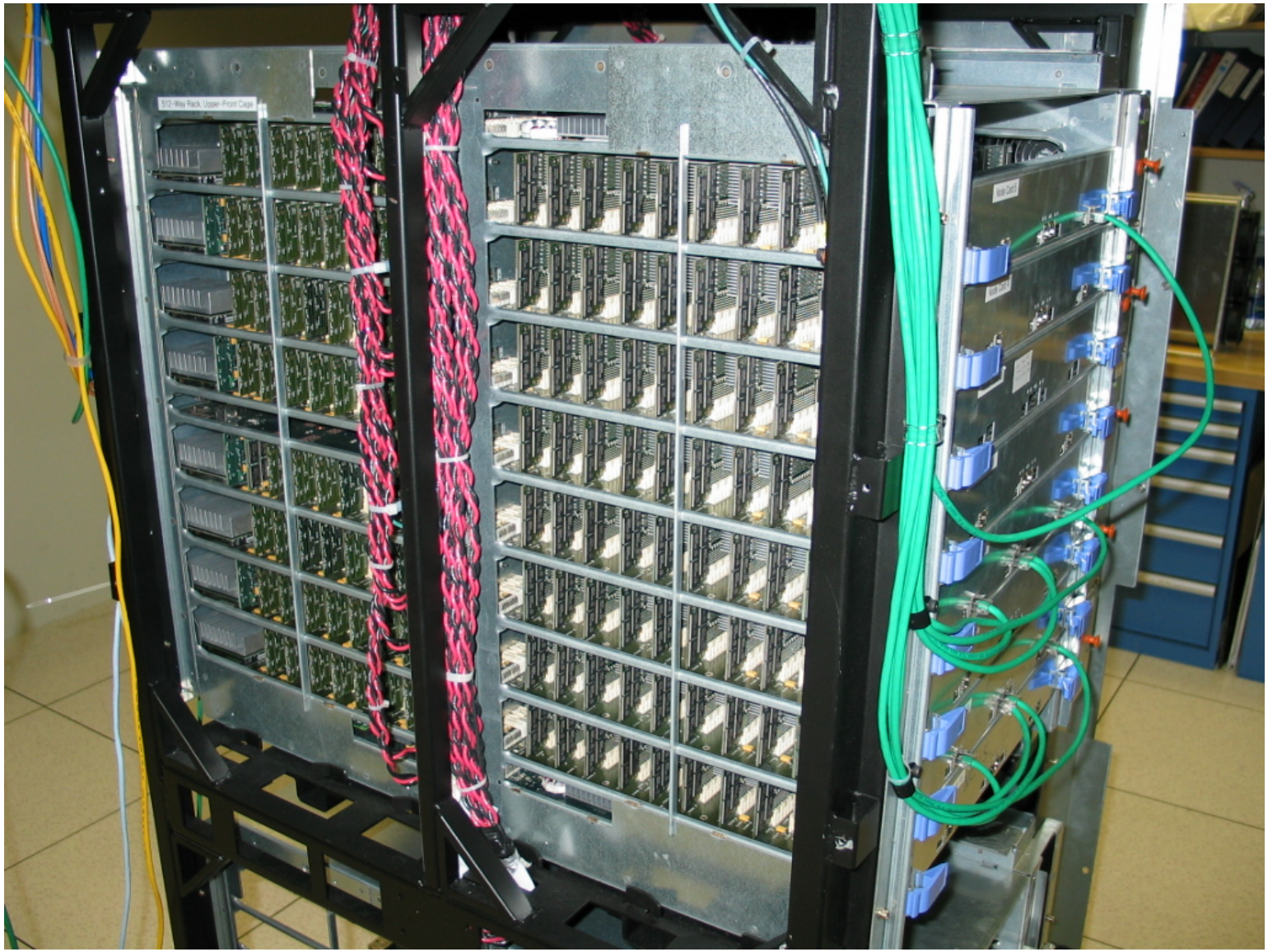
Dual Node Compute Card



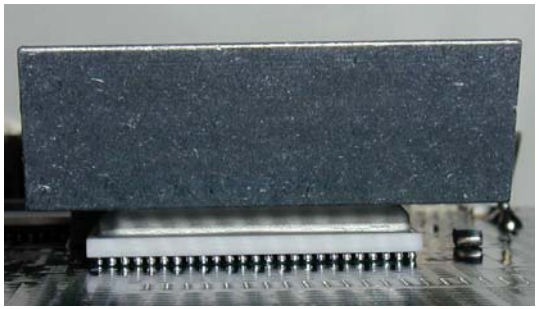

Midplane (450 pins) torus, tree, barrier, clock, Ethernet service port
16 compute cards
Ethernet-JTAG FPGA
dc-dc converters
2 optional IO cards
32- way (4x4x2) node card
IO Gb Ethernet connectors through tailstock
Latching and retention

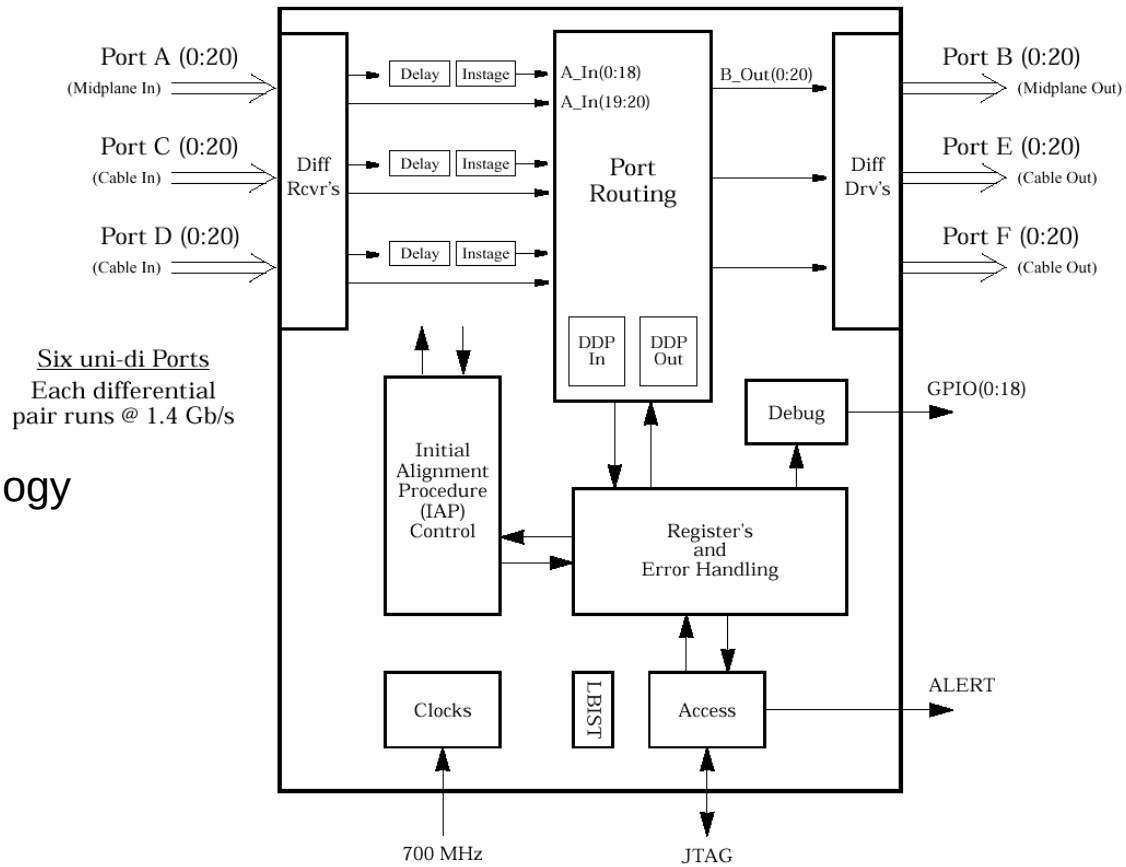

¼ of BG/L midplane (128 nodes)
Compute cards
IO card
dc-dc converter

512 Way BG/L Prototype



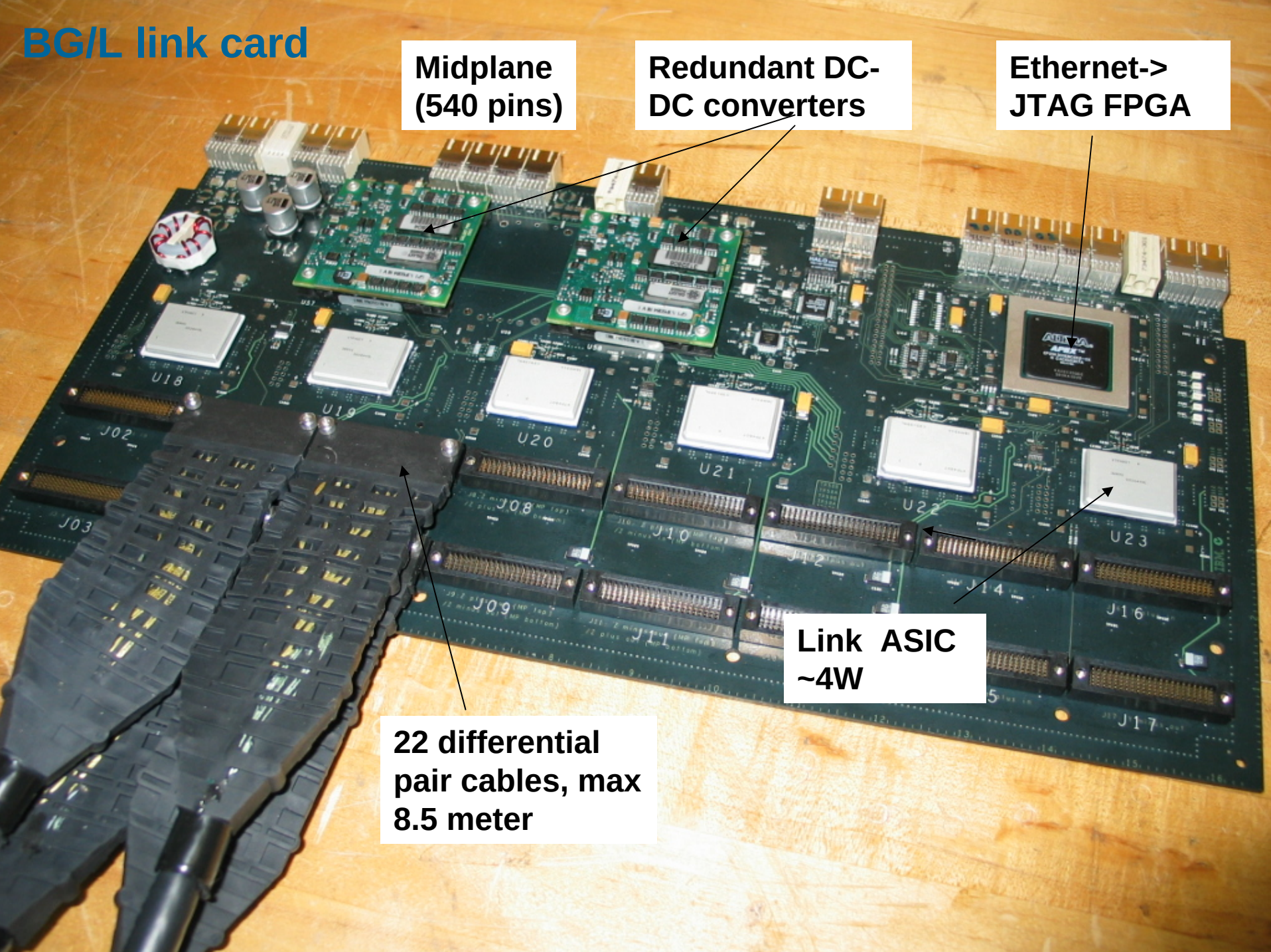
BlueGene/L Link Chip



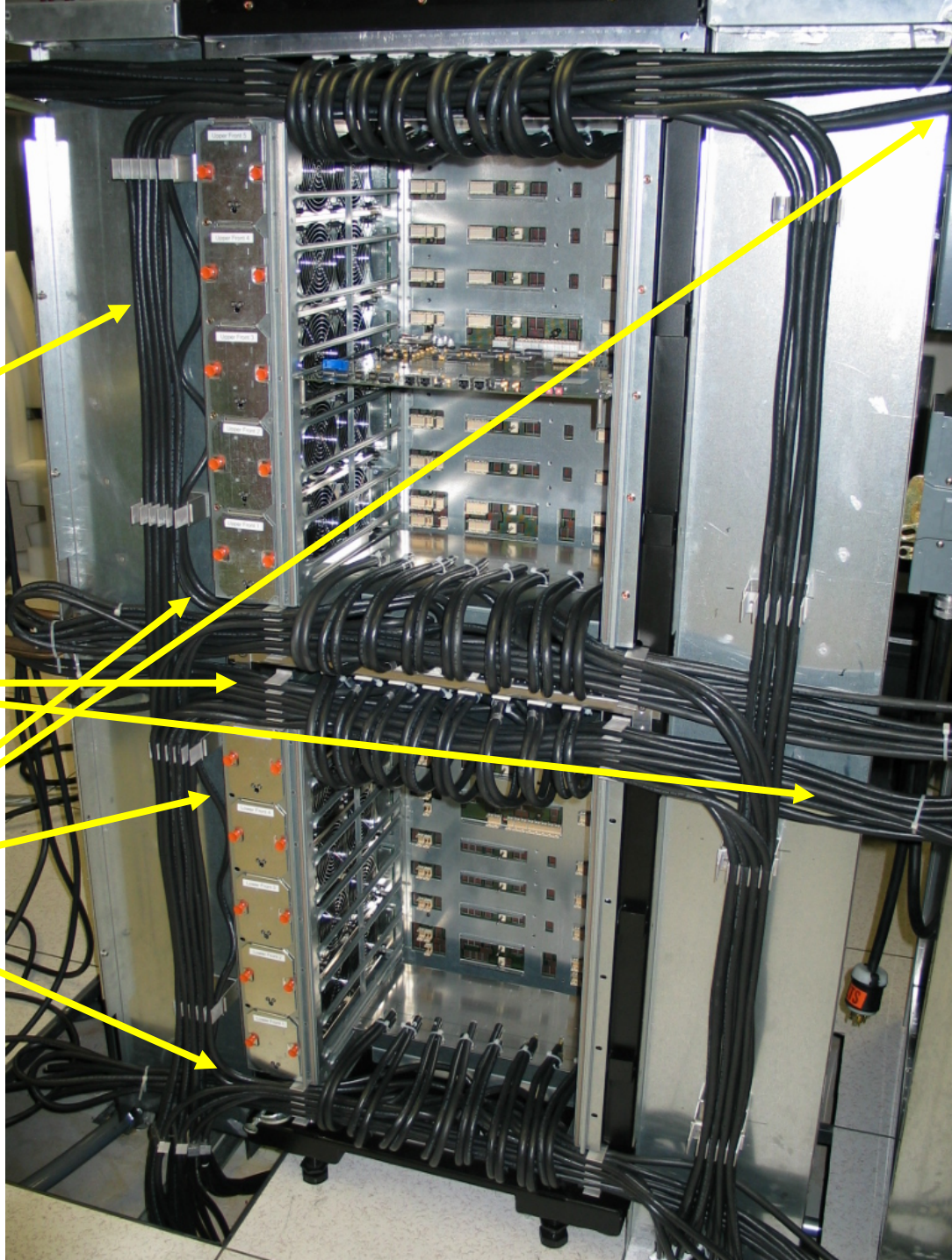


- IBM CU-11, 0.13 µm technology
- 6.6 mm die size
- 25 x 32 mm CBGA
- 474 pins, 312 signal
- 1.5 Volt

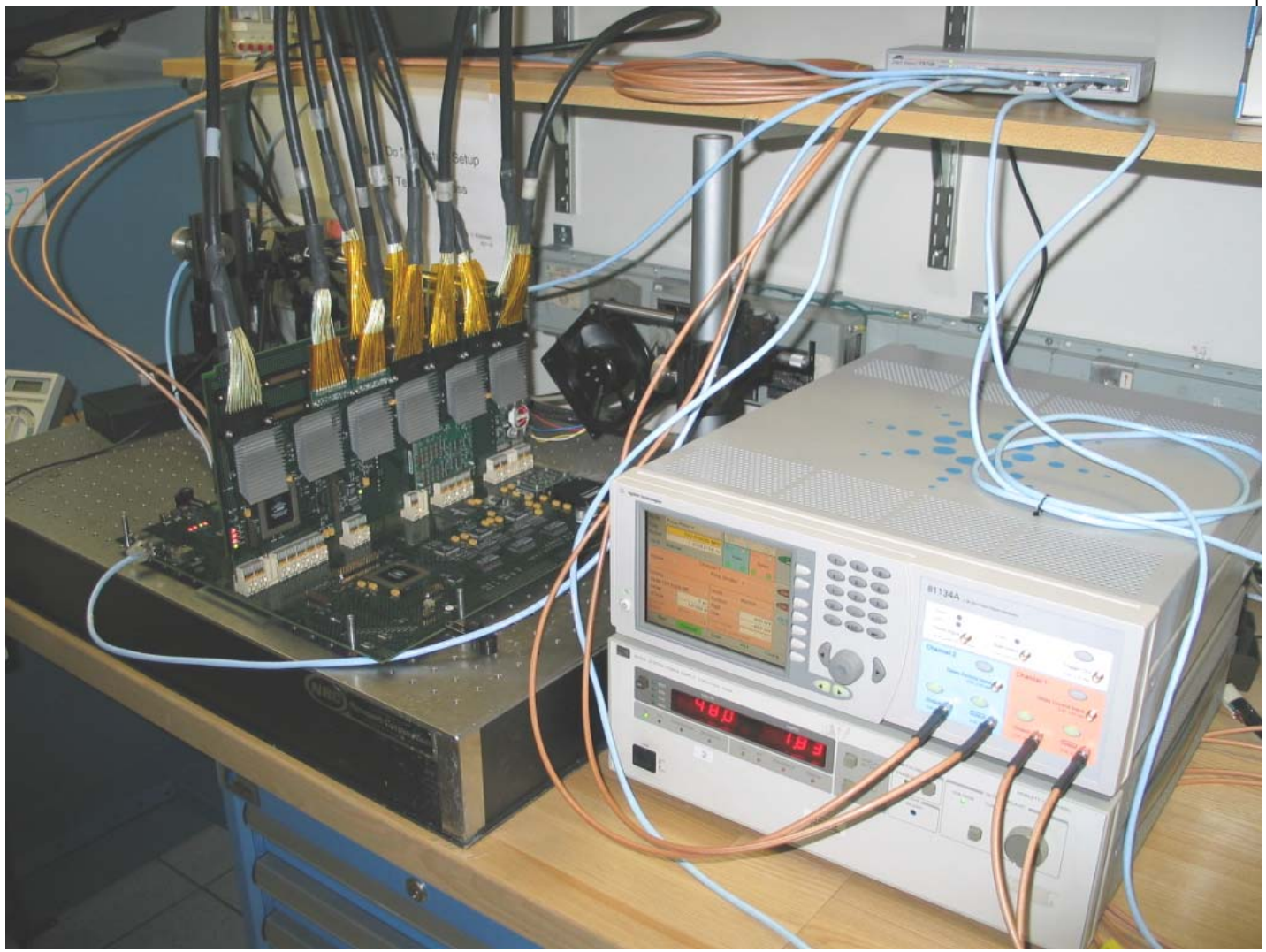
BG/L link card



BG/L rack, cabled

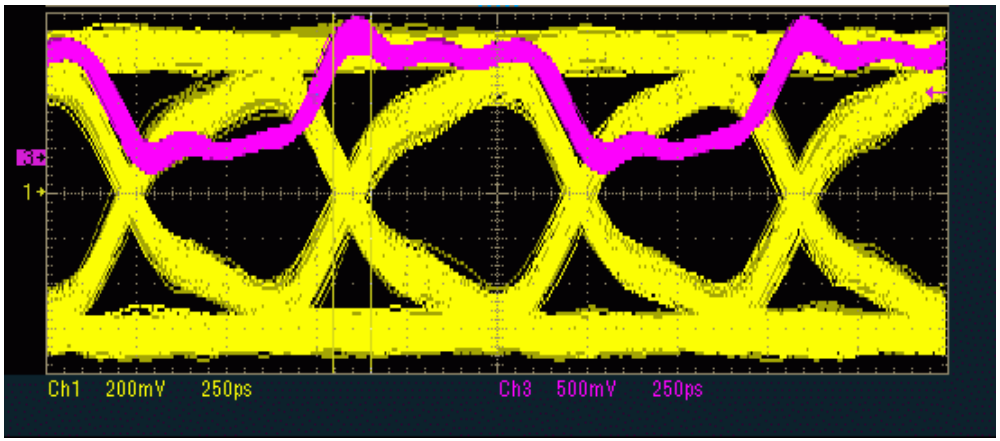


BLL Measurement Setup

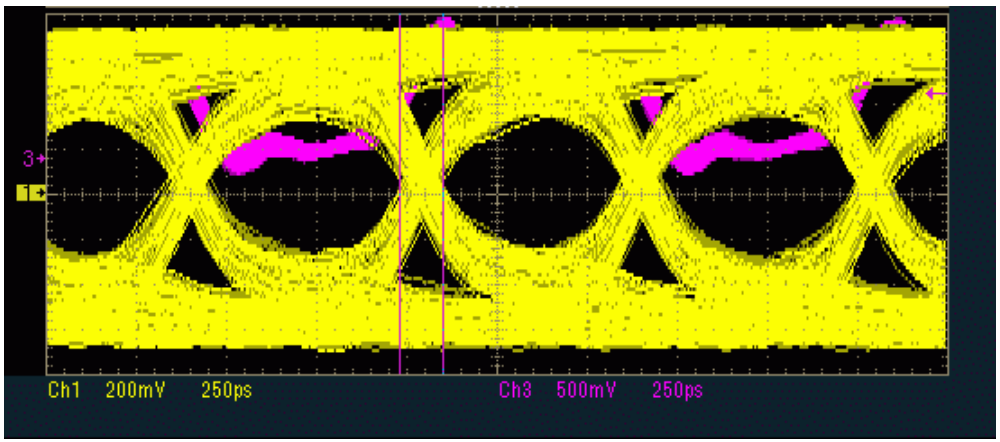


BlueGene/L Link “Eye” Measurements

1.6 Gb/s

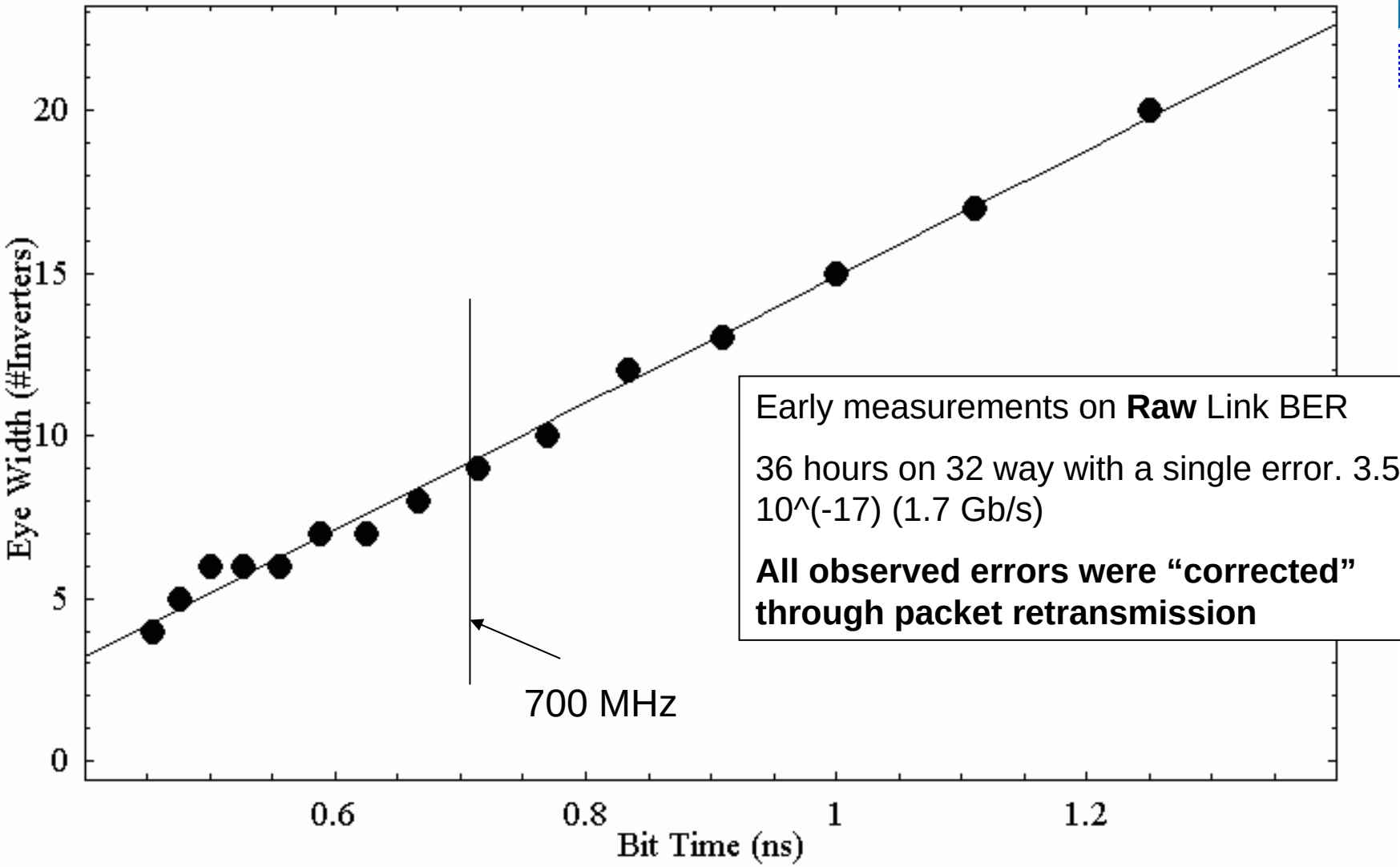


Signal path includes module, card wire (86 cm), and card edge connectors



Signal path includes module, card wire (2 x 10 cm), cable connectors, and 8 m cable

Link Performance Exceeds Design Target



Early measurements on **Raw** Link BER

36 hours on 32 way with a single error. 3.5 10^(-17) (1.7 Gb/s)

All observed errors were "corrected" through packet retransmission

Bit Error Rate Measurements

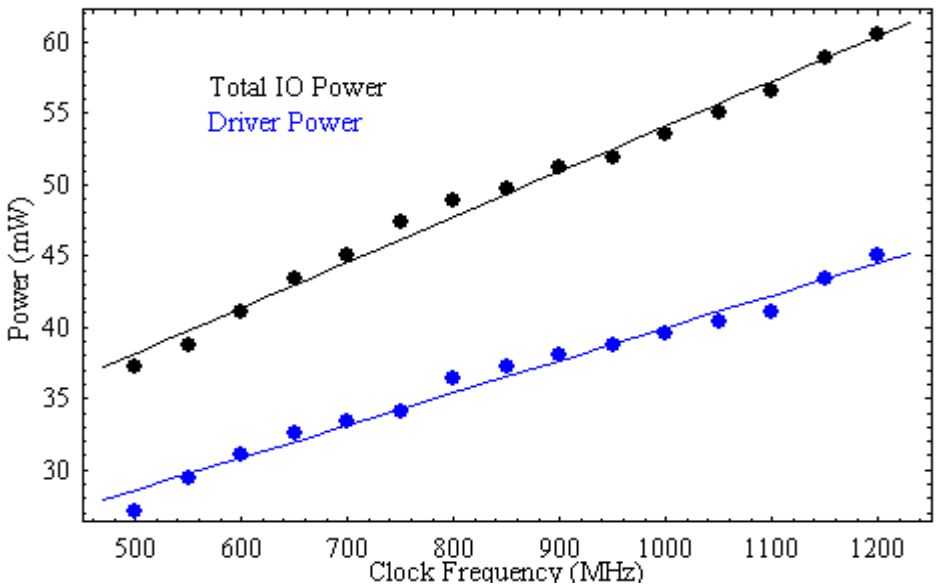
- Average data rate for experiment exceeds 260 Gb/s with 24% of bits transmitted through 8-10 m cables
- In over 4900 total hours of operation over 4.6 x 10^{18} bits have been transferred with only 8 errors observed (one error through 8-10 m cables)
- All errors were single bit (detectable by CRC)
- Aggregate midplane BW=8.4 Tb/s, at BER of 10^{-18} we expect a single bit error about every 33 hours per midplane
- Based on these results, packet resends due to CRC detected link errors will not significantly degrade BG/L performance

Data Rate (Gb/s)	Time (hours)	Total bits	Err	BER
1.4	335	2.3 x 10^{17}	0	4.4 x 10^{-18}
1.5	184	1.3 x 10^{17}	0	7.5 x 10^{-18}
1.6	893	9.3 x 10^{17}	0	1.1 x 10^{-18}
1.7	2139	2.0 x 10^{18}	1	4.9 x 10^{-19}
1.8	607	6.3 x 10^{17}	6	9.6 x 10^{-18}
1.9	512	5.0 x 10^{17}	0	2.0 x 10^{-18}
2.0	289	2.2 x 10^{17}	1	4.5 x 10^{-18}
1.4-1.7	3551	3.3 x 10^{18}	1	3.0 x 10^{-19}
1.8-2.0	1408	1.4 x 10^{18}	7	5.1 x 10^{-18}
Total	4959	4.7 x 10^{18}	8	8.9 x 10^{-19}

BER test status: 6/9/03

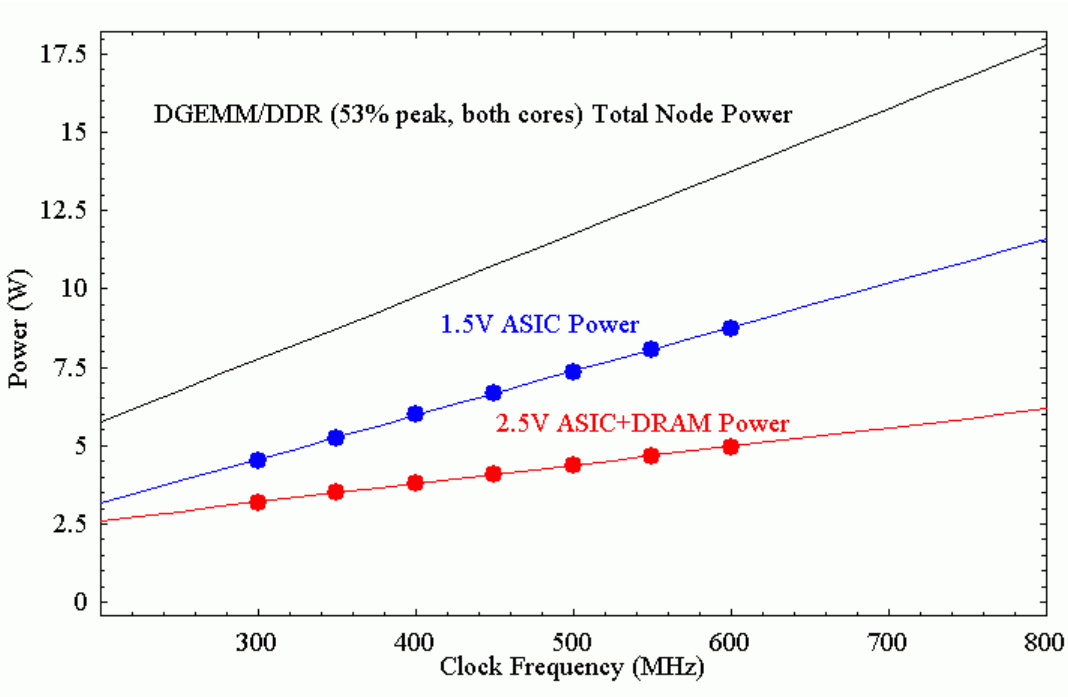
Link I/O Power per Bit

- Low power per bit is a key design feature due to the large number of high speed links in the BG/L torus and tree networks
- Measured power per bit is in excellent agreement with simulation



Note: Data rate is 2X clock frequency

BlueGene/L Compute Node Power

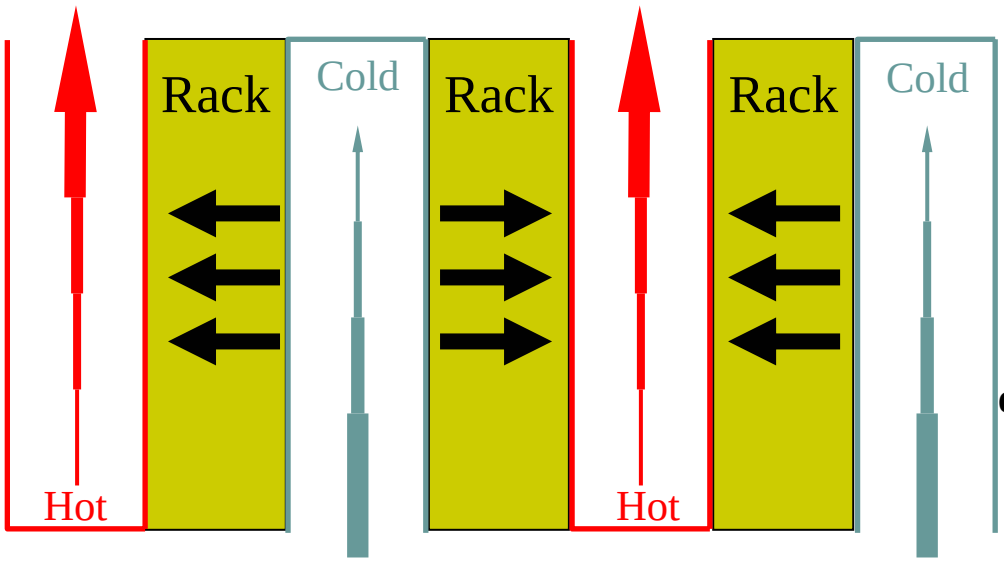


Power (W) for various programs	500 MHz	700 MHz
DGEMM/DDR-ASIC	8.7	11.5
DGEMM/DDR-Mem	3.1	4.3
DGEMM/DDR-Total	**11.8**	**15.8**
DGEMM/L3-ASIC	9.2	12.1
DGEMM/L3-Mem	1.6	1.6
DGEMM/L3-Total	**10.8**	**13.7**
MemXfer-ASIC	6.4	8.2
MemXfer-Mem	3.7	5.0
MemXfer-Total	**10.1**	**13.2**

BlueGene/L 512 Way Prototype Power

Maximum Power (W)		500 MHz		700 MHz	
Unit	**Num**	**Unit Pwr**	**Total Pwr**	**Unit Pwr**	**Total Pwr**
Node Cards	16	390	6240	519	8304
Link Cards	4	21	84	26	104
Service Card	1	17	17	17	17
dc-dc Conversion Loss	---	---	791	---	1051
Fans	30	26	780	26	780
ac-dc Conversion Loss	---	---	950	---	1231
Midplane Total Power	---	---	8862	---	11487
64k System Power (kW)	128	8.862	1146	11.487	1470
MF/W (Peak)	---	---	231	---	250
MF/W (Sustained)	---	---	160	---	172

BG/L L<->R airflow, direct from raised floor



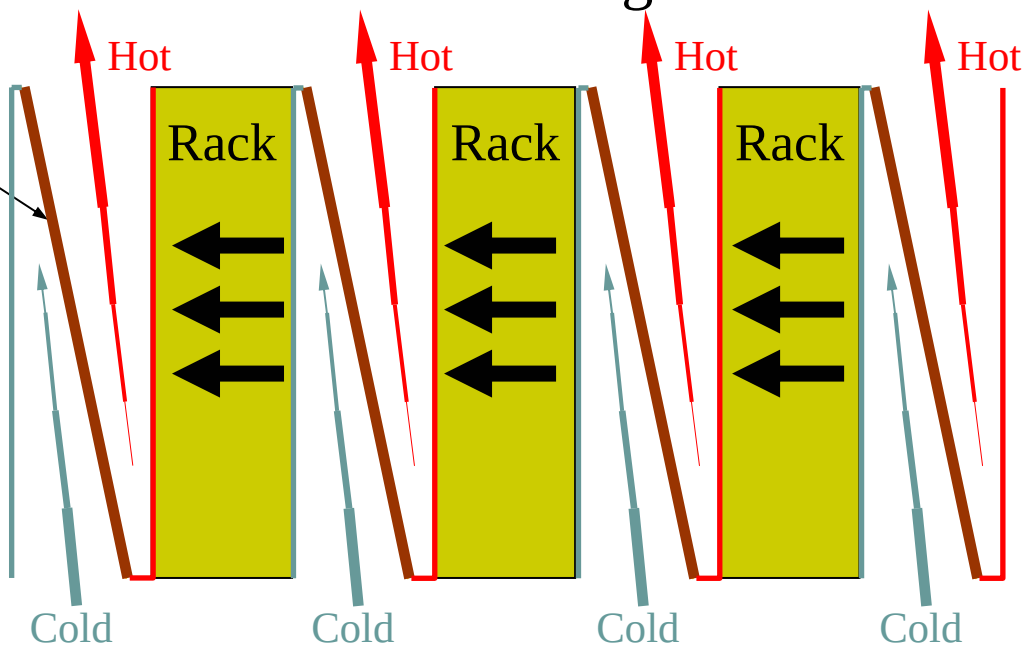
etc.

Flow rate in cold duct is largest at bottom; flow rate in hot duct is largest at top.

This scheme has same duct area, top to bottom, regardless of flow rate.

Alternative Ducting Scheme

Thermal-Insulating Baffle

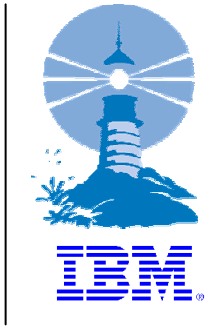


etc.

Alternative Ducting: Ducts are larger where flow is greater (Tj ~10C lower)

Shawn Hall 4-3-02
02-04-03 Angled Plenums

BG/L Reliability & Serviceability



- Redundant bulk supplies, power converters, fans, DRAM bits.
- ECC or parity/retry with sparing on most buses.
- Extensive data logging (voltage, temp, recoverable errors, …) and failure forecasting.
- Uncorrectable errors cause restart from checkpoint after repartitioning.
- Only fails early in global clock tree, or certain failures of link cards, require immediate service.

Summary

- Exploiting low power embedded processors, ASIC system-on-chip, and dense packaging enables large improvements in peak performance, cost/performance, floor space, and total power consumed over previous supercomputers.
- 512 way prototype is complete and all major functional subsystems are operational.
  - Compute and IO nodes with Gb Ethernet
  - Tree, torus and global interrupts
  - Control system
- Power and performance of half-rack 512 way prototype meet the design goals required to build a 64k node BG/L system.
- The success of BlueGene/L depends on the number and variety of applications that can be ported to run efficiently on the hardware.